Joint linear modeling of transcriptomics and proteomics reveals molecular coordination in cancer metastasis

Ronit Sharma*, Nikolaos Meimetis*, Arjana Begzati, Shashwat Depali Nagar, Benjamin Kellman, Hratch M. Baghdassarian
*contributed equally

Author Information

Author Affiliations

Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Nikolaos Meimetis, Hratch M. Baghdassarian

Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA, 92093, USA
Arjana Begzati

School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332 USA
Shashwat Depali Nagar

Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, 02139 USA
Benjamin Kellman

Unaffiliated
Ronit Sharma

Contributions

R.S.: Conceptualization, Transcriptomic Analysis, Writing N.M.: Conceptualization, Neural Network and Feature Analysis, Writing, and Editing A.B.: Feature Analysis, Writing S.D.N.: Power Analysis, Writing B.K.: Conceptualization H.M.B.: Conceptualization, Analysis, Writing, Editing, and Supervision.

Corresponding authors

Correspondence to Hratch M. Baghdassarian (hmbaghda@mit.edu).

Abstract

Cancer metastasis, a process in which cancer cells migrate to secondary sites, accounts for 90% of cancer deaths. While many machine learning models have been developed and applied to predict metastasis, they tend to be restricted to tumor types or classification tasks. Here, we apply a pan-cancer machine learning model that is easily interpretable and directly links the transcriptomic and proteomic profiles of hundreds of cancer cell lines with metastasis in a tissue-agnostic manner. We show that, for the relatively small sample size here, linear models perform just as well as non-linear models, including neural networks. We demonstrate that transcriptomics outperforms proteomics for predictive tasks due to more comprehensive genome coverage. However, the combination of both transcriptomics and proteomics leads to an overall increase in performance and we provide evidence indicating complementary information available between these modalities. Features identified as most important by the joint omics model are over-represented in biological pathways related to metastasis, suggesting that the model learns relevant information that may be useful in identifying universal biomarkers or therapeutic targets of cancer metastasis. Finally, we identify feature modules with synergistic or antagonistic effects that suggest potential mechanisms that amplify metastatic phenotypes.

Introduction

Cancer is the second-leading cause of death in the United States, and metastasis accounts for 90% of these deaths. During metastasis, cancer cells with increased motility detach from the primary tumor, migrate, and invade and adhere to organs. To enable this, metastatic cells undergo the epithelial-mesenchymal transition and develop specific properties such as resistance against anoikis and the formation of invadopodia. Elucidating the molecular mechanisms underlying these various properties that favor metastasis has motivated multiple therapeutic approaches to treat cancer[1–3]. As such, placing molecular features in the context of metastasis in a data-driven manner can reveal novel genes driving metastatic properties and has implications for therapeutic targets.

Cell motility is a fundamental property of metastasis, and it varies substantially across the tumor microenvironment and cell lines[4]. Metastasis changes according to contexts such as genetics, extracellular cues, and spatial location that alter the molecular state of cancer cells[5,6]. Consequently, the combinatorial space of molecular features underlying metastatic phenotypes is difficult to decipher[7], and it is difficult to identify generalizable mechanisms of metastasis.

Systems biology enables the study of such complex functions, with computational models helping to elucidate how molecular and phenotypic profiles are related[8–10]. In particular, machine learning (ML) approaches can leverage large omics datasets[11] to train predictive models that, for example, identify disease biomarkers, provide insights into gene regulation, and guide the designing of drugs[12]. Such approaches have been used specifically to study cancer metastasis. For example, migration rates have been predicted from feature sets that do not involve global

molecular profiles, such as cell morphology[13]. Additionally, there is extensive work on predicting metastasis from omics for tumors in specific tissues or cancer types[14]. While some pan-cancer approaches exist, they are focused on classification tasks, and some also use complex neural nets that are not easily interpretable[15,16].

Here, instead, we provide a prediction of metastatic potential as a continuous variable using an easily interpretable linear model. We leverage publically available transcriptomic, proteomic, and phenotypic data from DepMap[17,18] and MetMap[19], which spans 21 types of solid tumors, indicating that our approach is generalizable and, since we do not incorporate the tissue-specific information reported in MetMap, tissue-agnostic. Our analysis of transcriptomics measurements shows that, on these data, simple linear models can perform as well as more complex, state-of-the-art non-linear ML approaches. Additionally, while transcriptomics seems to be the more informative modality in this case due to wider coverage, we observed that the combination of both transcriptomics and proteomics increases predictive performance. Finally, we demonstrate that specific features from these modalities are combined by the model in a complementary manner, with priority going to those associated with biological pathways relevant to metastasis. Computational simulations also identified non-additive interactions amongst modules of genes that amplified metastatic phenotypes despite not being amongst those top-ranked by the linear model. Interactions converged on hub proteins regulating key signaling pathways that influence metastasis such as MAPK and Wnt. Altogether, this study demonstrates the utility of using multi-omics data in interpretable models not only for predictive performance, but to understand how features coordinate across modalities in relation to a phenotype of interest.

Results

Linear Models Perform At Least as Well as Non-Linear Models in Predicting Metastatic Potential from Transcriptomics

In order to determine whether the global abundance of gene products is informative of metastasis, we leveraged DepMap's cell line measurements[17]. Specifically, we fit a dataset containing 481 samples on a set of machine learning models that take as input protein-coding genes measured by RNA-sequencing and predict as output metastatic potential reported by MetMap 500[19]. After hyperparameter tuning and rank-ordering of 9 linear and non-linear machine learning models by mean performance (Fig. S1a, see Methods for Details), we further assessed the top three models (support vector regression (SVR) with radial basis function, polynomial, and linear kernel) using consensus hyperparameters, alongside single and ensemble neural nets. Despite potentially complex relationships between molecular features and cell phenotypes such as cancer cell metastasis, we observed no significant differences (Mann-Whitney U, Benjamini-Hochberg FDR correction, $q \geq 0.85$) between the linear SVR and the non-linear models, including neural nets (Fig. 1a, Fig. S4a). These results are further supported by the lack of significant differences ($q \geq 0.86$) between all 9 model predictions during hyperparameter tuning using nested CV (Fig, S1a, Table 1).

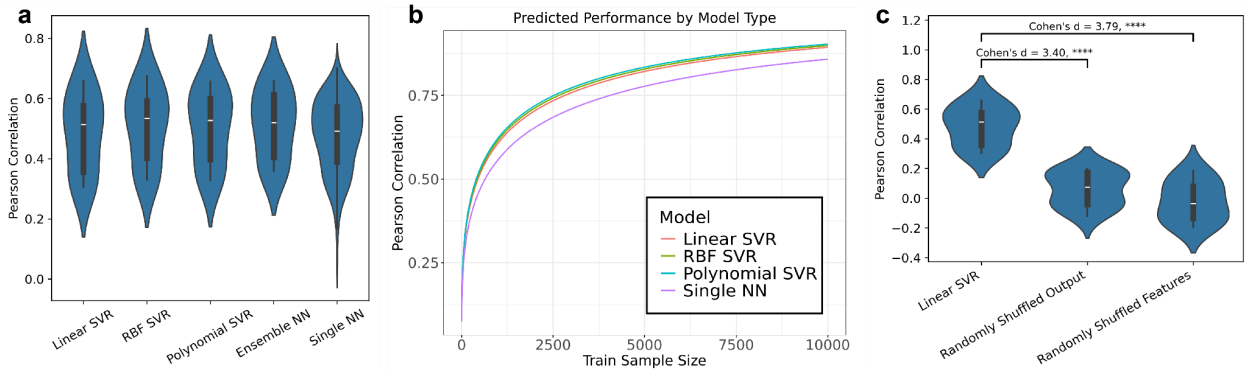

Fig. 1: Linear models exhibit predictive power for metastasis from transcriptomics. *All statistical significance is determined using a Mann-Whitney U (MWU) test and the FDR is controlled for multiple testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Models' predictive performance is assessed by Pearson correlation. Violin plots visualize performance distribution output from 10-fold CV.* ***a*** *Comparison of consensus linear SVR model to other consensus SVR models' and neural nets' (x-axis) performance (y-axis) (two-sided MWU test). No comparisons have $q \leq 0.1$.* ***b*** *Estimates of Pearson correlation (y-axis) for each consensus model type across training sample size (x-axis) from 1 to 10,000 training samples. Estimates are derived from a power function regression modeling Fisher z-transformed Pearson correlation as a function of Train Sample Size and including model type as a covariate.* ***c*** *Consensus Linear SVR model performance (y-axis) compared to random baselines (x-axis) (one-sided MWU test).*

Because the number of features used is much larger than the number of available samples, and because more complex models with many parameters such as neural nets require large sample sizes, we conducted a power analysis to determine how each model's performance changes with sample size (Fig. S2, see Methods for details). First, we determined that the change in performance across sample size is best explained by a power function, as opposed to a linear or exponential function (Table 2). Next, we fit a power function regressing model performance against the training sample size, and we included the model type as a covariate to assess differences between them (Fig. S3). When controlling for sample size, this regression estimates the nonlinear SVRs have a slightly ($10^{0.01}$ = 1.02-fold) higher Fisher z-transformed Pearson correlation than the linear SVR. However, practically speaking, this difference is negligible (Fig. 1b). For example, when extrapolating to 10,000 samples, all model types are estimated to have ~0.90 Pearson correlation. Furthermore, the single neural nets are estimated to perform more poorly than the linear SVR; the Fisher z-transformed Pearson correlation is ($10^{-.05}$ = 0.9 times) lower than the linear SVR, predicting ~0.86 Pearson correlation at 10,000 samples. Results for MSE are qualitatively similar (Fig. S3). This indicates that the sample size is a much more substantial limiting factor in model performance than model type, and that linear models may continue to perform as well as non-linear models even at larger sample sizes.

Since support vector regression (SVR) models were the top performers (Fig. S1a), and because non-linear kernels did not outperform the more interpretable linear kernel, we proceed with downstream analyses using the linear SVR. With a mean 10-fold CV Pearson correlation of 0.48, we confirm that the consensus linear SVR model outperforms random baselines in which the output values or input features were randomly permuted prior to fitting the model with the

same hyperparameters (Fig. 1c, Fig. S4b). Altogether, these results indicate that machine learning models can explain ~22% of the variance (quantitated by the mean 10-fold CV coefficient of variance of the consensus linear SVR) in metastatic potential from transcriptomics, and that, for these data, a more interpretable linear model suffices in capturing this relationship.

Transcriptomics and Proteomics Demonstrate Predictive and Informational Discrepancies with Respect to Metastasis

Next, because protein-level gene abundance may contain additional information about cell phenotypes independent of that available in RNA-level gene abundance[20], we asked how models may perform on proteomics data. We again identified consensus models for the top 3 performers using proteomics (Ridge and SVRs) (Fig. S1b). Because the proteomics data had fewer samples, we compared it to the transcriptomics consensus linear SVR model run on an equal sample size from the power analysis. Using 10-fold CV, we found that the transcriptomics model outperformed the proteomics models as measured by MSE (Fig. 2a, Fig. S5a). Additionally, proteomics models did not outperform the transcriptomics one or performed worse (in the analogous linear SVR) than it as measured by Pearson correlation, and in some instances, did not outperform random baselines (Fig. S6). For additional details and discussion regarding these comparisons, see Supplementary Results - Transcriptomic and Proteomic Comparison Using All Features.

Next, we asked whether transcriptomics may outperform proteomics due to the fact that it has more comprehensive coverage of the genome (19,138 transcriptomic features as compared to 10,969 proteomic features). We thought this may be the case given that the feature selection step of the hyperparameter tuning pipeline tended to select much fewer proteomic features, indicating that when the relevant features are measured, proteomics can get more information from those features as compared to transcriptomics. By comparing the linear SVR performance on either modality fit on only the intersection of features, we saw that there was no longer a significant difference in performance as measured by MSE between transcriptomics and proteomics (Fig. 2b), as well as a much smaller effect size (Cohen'd D of -0.44 for the intersection of features and -1.76 for all features) . Given that the absence of informative features in proteomics measurements could explain the decreased performance of these models, we also checked the extent of overlap between proteomic features and highly-ranked features in the consensus linear SVR transcriptomic model. We observed that within the top 100 features ranked by the transcriptomic model, a substantial fraction (31%) were not present in the proteomic dataset (Fig. 2c). Altogether, these results indicate that the better performance of transcriptomics over proteomics can be explained by the fact that it measures more features, including a subset informative of metastasis; however, when relevant proteomic features are measured, they are highly informative. For additional details and discussion regarding these comparisons, see Supplementary Results - Transcriptomic and Proteomic Comparison Downsampling Features.

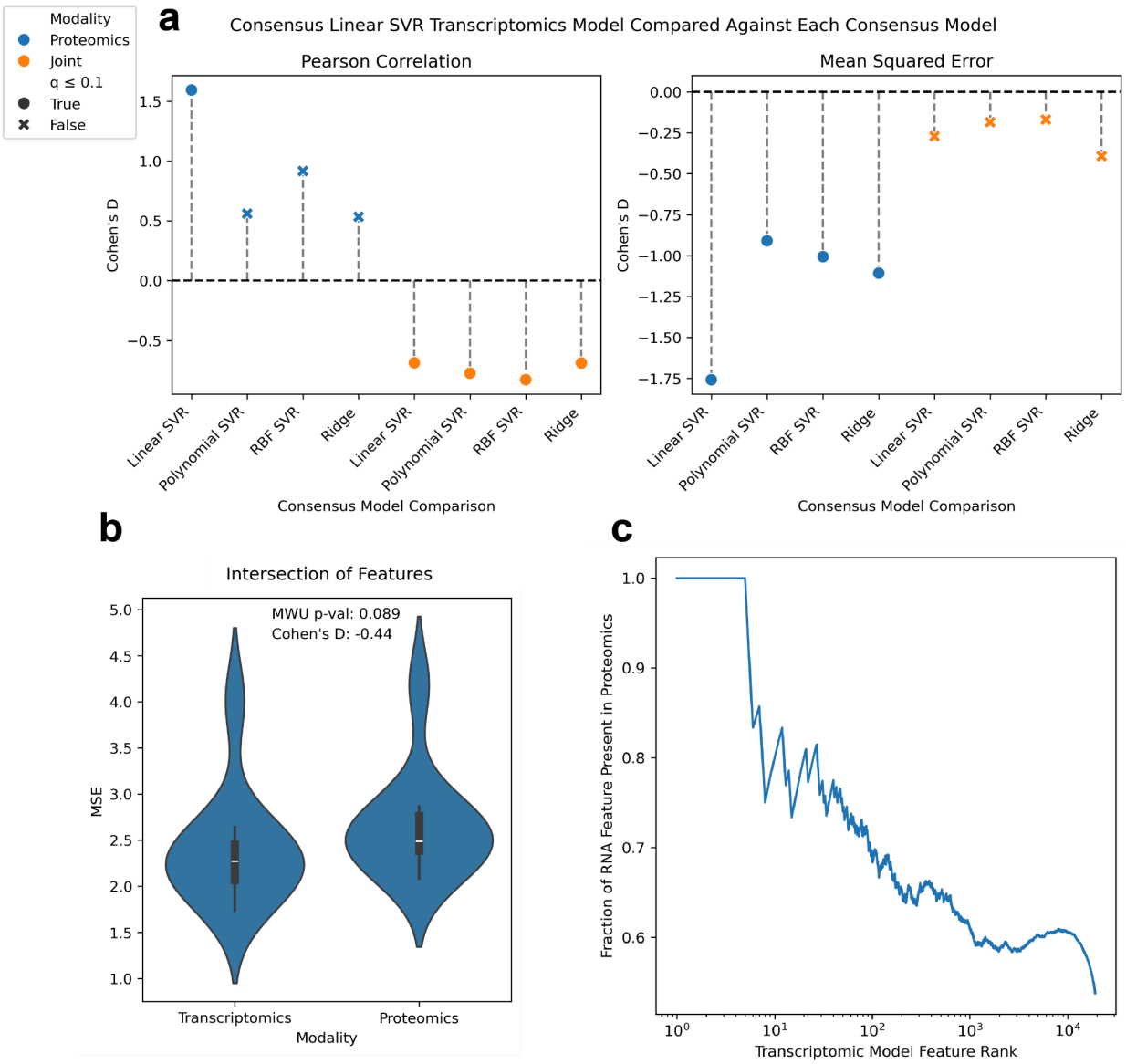

Fig. 2: Proteomics and Transcriptomics Offers Distinct Views of Metastasis. a *Lollipop plots visualize comparisons between top consensus models (x-axis) identified in proteomics (blue) or joint omics (orange) with respect to the consensus transcriptomic linear SVR. Models' predictive performance is assessed by Pearson correlation (left panel) and mean squared error (right panel). Comparison effect sizes are quantified using Cohen's d (negative values mean the effect was larger in the transcriptomic model and positive values mean the effect was smaller in the transcriptomic model). Significance (two-sided Mann-Whitney U test, Benjamini-Hochberg FDR correction, q ≤ 0.1) is indicated by the scatter point shape. Performance for each consensus model (x-axis) is assessed by 10-fold CV, and compared against the linear SVR performance at the same sample size using the power analysis previously described.* ***b*** *Violin plots visualize distribution of 10-fold CV performance measured by MSE comparing linear SVR performance fit on transcriptomics or proteomics (x-axis). Model fitting and assessment was conducted on the intersection of both samples and features between proteomics and transcriptomics to make them as directly comparable as possible.* ***c*** *Line Plot of the fraction of transcriptomic features that overlaps with proteomic features (y-axis) as a function of the number of transcriptomic features at a given rank-order (x-axis). Feature ranks are obtained from fitting the linear SVR model with consensus*

hyperparameters on the entire transcriptomic dataset and ordering by absolute value of the model coefficient.

Given the discrepancies between these modalities, we asked whether incorporating both transcriptomics and proteomics could enhance models' predictive performance. After again identifying top-performing consensus models (Fig. S1c), we found that the joint omics models outperformed the consensus linear SVR transcriptomics model as measured by Pearson correlation, with medium to large effect sizes (-0.69 ≥ Cohen's d ≥ -0.83) (Fig. 2a, Fig. S5b). The best consensus model for joint omics (SVR with RBF kernel) had a mean Pearson correlation and MSE of 0.51 and 2.28, respectively, as compared to the consensus linear SVR for transcriptomics at the same sample size having a Pearson correlation and MSE of 0.44 and 2.21, respectively. Pearson correlations for all consensus joint omics models ($0.50 \geq r \geq 0.51$) are comparable to that of the consensus transcriptomic model assessed on the entire transcriptomic dataset ($r = 0.48$), despite only having approximately half the samples available. Finally, all consensus joint omics models performed better than random baselines (Fig. S7).

However, joint omics models did not outperform transcriptomics as measured by MSE, and the transcriptomics linear SVR model had lower standard deviations (0.09 Pearson correlation, 0.43 MSE) than consensus joint omics models (minimum across consensus models of 0.14 Pearson correlation, 0.79 MSE). Altogether, given that hyperparameter tuning was performed using MSE, the orthogonal Pearson correlation metric indicates that joint omics offer a substantial boost in predictive performance.

Multi-omic Modeling Captures Metastasis-Associated Coordination Between Modalities

Since feature selection and model performance differences suggested that proteomic and transcriptomic models captured distinct metastasis-related signals, we reasoned that the joint omics model could help contextualize these features in relation to phenotype. We proceeded in this analysis with the consensus linear SVR as it had the lowest standard deviation in performance for both metrics during 10-fold CV, and was not significantly outperformed by the other consensus models, including the less interpretable non-linear models (Fig. S8b). We rank-ordered features by the absolute value of their coefficient in the model fit to the full dataset and ran an over-representation analysis (ORA) of the top 500 features, separated by those with negative or positive coefficients (Fig. S9). Many of the overrepresented gene sets are related to metastasis. For example, positive regulation of cell migration is any process that activates or increases cell migration, a key feature of metastatic cancer cells. Additionally, chemotactic migration (chemotaxis) is required by cancer cells to invade the surrounding tissues of the tumor, with EGF and other growth factors serving as factors driving cancer cells to respond to this stimulus and migrate (cellular response to growth factor stimulus)[21]. Integrins play a key role in cellular adhesion, as well as intracellular signaling pathways that regulate cytoskeletal organization and cellular death, all of which are strongly associated with cancer metastasis[22]. Finally, Reactive Oxygen Species (ROS) can trigger signaling pathways that lead to cellular migration[23].

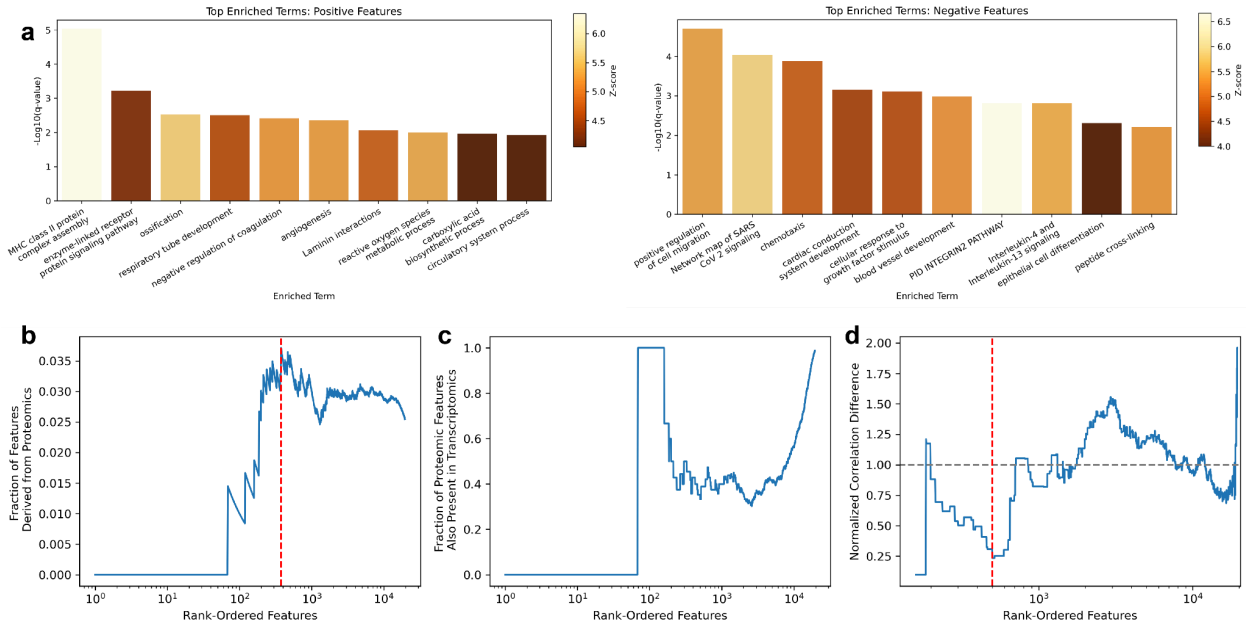

Fig. 3: Proteomics and transcriptomic features complement each other in driving metastatic programs. a *Barplots of the top 10 Metascape over-represented summary terms for features with negative (left panel) and positive (right panel) coefficients in the consensus linear SVR. Terms are rank-ordered by* $\log_{10}$*(q-value) (y-axis) and colored by z-score. Z-scores are aggregated by their average across individual members of each summary term.* ***b*** *Line Plot of the fraction of total features derived from proteomics (y-axis) as a function of the number of features at a given rank-order (x-axis). Features are rank-ordered by the absolute value of the model coefficient. The red dashed line represents the rank at which the proteomic feature fraction is maximized.* ***c*** *Line Plot of the fraction of selected proteomic features (y-axis) that overlap with selected transcriptomic ones as a function of the number of features at a given rank-order (x-axis). For a proteomic feature at a given feature rank, we define a corresponding overlapping transcriptomic feature as one that has a shared gene product and a feature rank at or below ("higher rank" - more important) that of the proteomic feature. Conversely, a non-overlapping transcriptomic feature is one that has a shared gene product with the proteomic feature and a feature rank above ("lower rank" - less important) that of the proteomic feature.* ***d*** *Line Plot comparing differences in proteomic-to-transcriptomic feature correlation for overlapping or non-overlapping genes (y-axis) at a given rank-order (x-axis). The correlation difference at a given feature rank is* $r_{\triangle} = r_{overlapping} - r_{non\text{-}overlapping}$*.* $r_{overlapping}$ *is calculated by identifying all proteomic features at a given rank that have overlapping transcriptomic features, taking the correlation between each feature pair across samples, and taking the average across pairs.* $r_{non\text{-}overlapping}$ *is similarly calculated, but instead identifying all proteomic features at a given rank that do not have overlapping transcriptomic features. We normalized to the mean* $r_{\triangle}$ *across all feature ranks. Thus, a value of 1 represents the average* $r_{\triangle}$ *across all features after normalization (horizontal gray line). The top 500 features are demarcated by the red dashed line.*

Next, we asked how the proteomic and transcriptomic data modalities were combined with regards to the model fit. During hyperparameter tuning, whereas all 19,138 features were selected for transcriptomics, only 500 of the 10,969 proteomic features were selected, representing 2.55% of all features used by the model. The selection of only a small subset of proteomic features but all transcriptomics features was consistent across model types (Table 3). However, proteomic features represent 3.40% of the top 500 features (odds ratio 1.36, Fisher's exact test one-sided p-value 0.14) and are maximized at the 384th rank at 3.71% (odds ratio

1.49, Fisher's exact test one-sided p-value 0.10) of the feature set (Fig. 3b). While over-representation in the top 500 features is not significant ($p > 0.05$), it is trending towards significance (Fig. S10). These results indicate that while proteomics comprise only a small subset of the features, they represent important additional information for the model. This aligns with our previous observations regarding the boost in performance provided by using the joint omics data (Fig. 2a, Fig. S5b) and the observed discrepancies between transcriptomic model rank and proteomic feature availability (Fig. 2c).

Next, given the variance of within-gene correlation between mRNA and protein abundance[20,24,25], we asked the extent to which gene products overlapped between these two modalities in the context of the joint omics model fit. While there was nearly complete overlap across all selected features, only 35.3% of proteomic features overlapped with transcriptomic features in the top 500, indicating a significant depletion (odds ratio 0.02, Fisher's exact test one-sided p-value 7.77e-11) of gene product overlap between mRNA and protein (Fig. 3c). This indicates that the model is accounting for redundant information between mRNA and protein.

Consequently, we set out to assess whether complementarity between mRNA and protein information is accounted for by the model fit. In instances where a proteomic feature was selected over its corresponding transcriptomic feature with a shared gene product at a given feature rank (non-overlapping), these may carry redundant information. This could indicate that the protein feature is more informative of cancer cell metastasis than the corresponding transcriptomic feature. In instances where both the proteomic feature and its corresponding transcriptomic feature with a shared gene product were included at a given feature rank (overlapping), there may be complementary information between the two modalities.[20] To explicitly test this, we assessed how the Pearson correlation between pairs of features differed between overlapping and non-overlapping features (Fig. 3d). We use higher Pearson correlation as a proxy for higher redundancy and lower Pearson correlation as a proxy for higher complementarity. While overlapping features almost always had higher Pearson correlation at a given rank (Fig. S11), the difference in Pearson correlation between overlapping and non-overlapping features tends to be lower than average at more important feature ranks. Similarly, it tends to be higher than average or close to average at more lowly ranked features (Fig. 3d). This supports the notion that, amongst highly ranked features, the model selects overlapping features when they are complementary and non-overlapping ones when they are redundant.

Low-ranked features exhibit non-additive interactions that strongly modulate metastatic potential

Since linear SVRs cannot capture non-additive interactions between features, we sought to identify feature interactions (see Methods for details). While top-ranked features did not exhibit significant interactions, we reasoned that synergistic or antagonistic relationships could emerge from co-dependencies among features with minimal individual effects but substantial combined influence. Using a genetic algorithm (GA) that optimizes the additional information gained from considering a feature interaction to identify modules of highly interacting features, we found 31

“high-confidence” interactions involving 28 features. These 28 features had an average feature rank of 9514 in the consensus linear SVR joint omics model and 15 features participated in more than one interaction (Table 4 sheet 2-3).

Given that the GA is optimized for features involved in modules of interactions (i.e., multiple recurring interactions amongst a core set of features), we hypothesized that these features may function as “hub” proteins in a protein–protein interaction (PPI) network[26]. To test this, we constructed a prior knowledge PPI network from the Omnipath[27] database, and we found that the features participating in high-confidence interactions had a significantly higher centrality (one-sided MWU p-val ≤ 2.83e-5) (Fig. 4a) than the rest of the features in the model, supporting their potential role as network hubs. Of the features participating in high-confidence interactions, the Spearman correlation between the number of interactions a feature participated in (ranging between 1 and 5, Table 4 sheet 3) and its eigenvector centrality was 0.51.

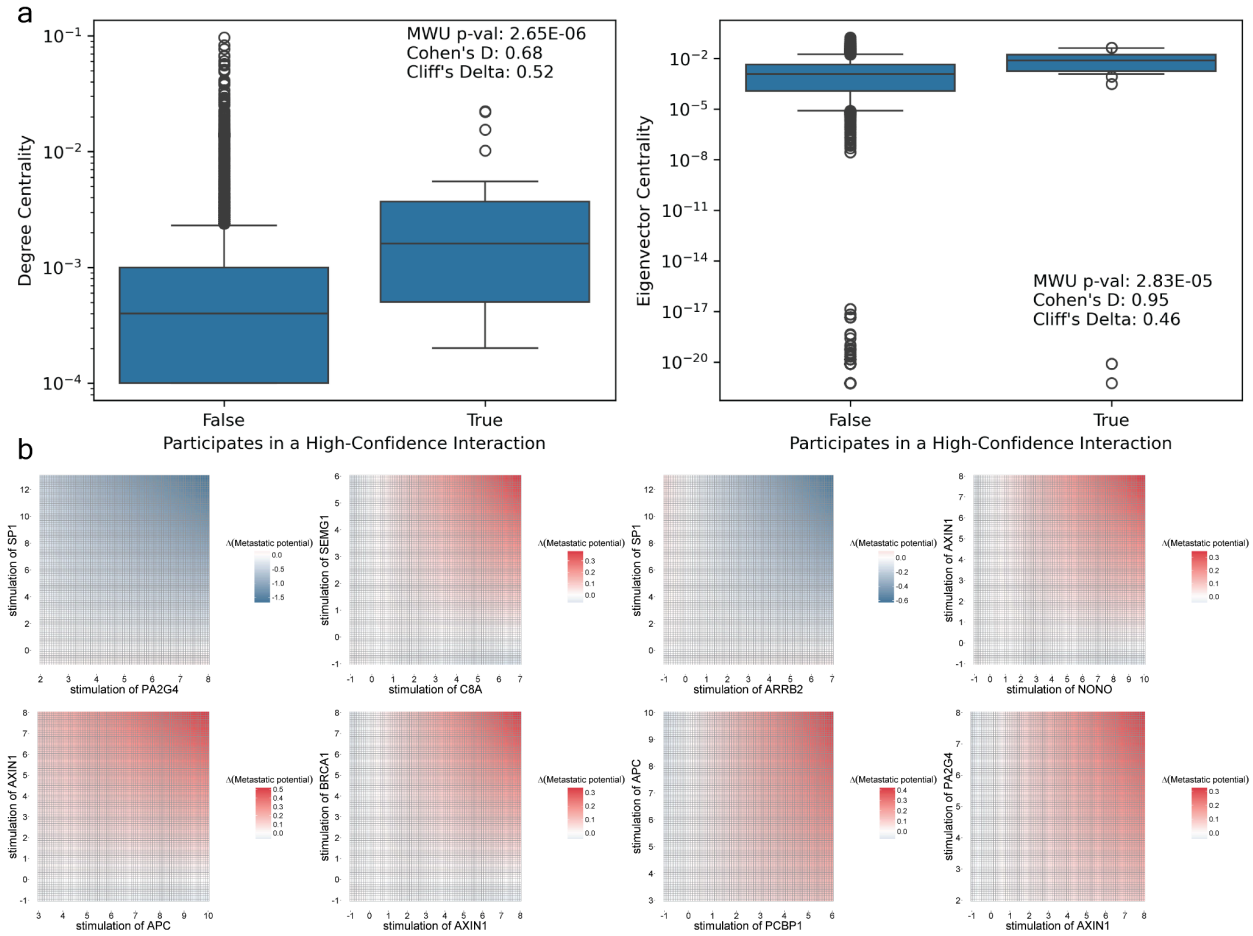

***Fig. 4: Recurring feature modules demonstrate consistent synergies and antagonisms**. **a** Boxplots visualizing the degree (left panel) and eigenvector (right panel) centrality (y-axis) for features in the consensus joint omics linear SVR model that were identified to participate in high-confidence interactions or not (x-axis). Each panel is annotated with the Mann-Whitney U p-value, Cohen’s D, and Cliff’s delta for its respective comparison. **b** Heatmaps of the top 8 high-confidence interactions ranked by the absolute value of the maximum effect. Effects are quantified as the change in predicted metastatic potential when the interaction is accounted for; regions of negative differences indicate antagonistic interactions (blue)*

and those of positive differences indicate synergistic interactions (red). As a reference for the relative change, empirical metastatic potential values range between -4.39 and 2.78.

To better understand the synergistic or antagonistic effects of these features on metastatic potential, we used the interaction model to compare differences in metastatic potential when accounting for interactions. By fitting these models on the complete dataset, and restricting *in-silico* perturbations near the empirically observed expression range of each feature, we keep every simulation mostly interpolative, resulting in highly reliable predictions of metastatic potential (Fig. S13, see Methods for details). The top 8 interactions by absolute value of maximum effect repeatedly included SP1, AXIN1, APC, and PA2G4. Three of these four features had consistent directional effects regardless of their interacting partner (SP1 is antagonistic, APC and AXIN1 are synergistic) (Fig. 4b). APC and AXIN1 both act as antagonists of Wnt/β-catenin signaling, a major contributor to tumorigenesis that controls proliferation and metastasis[28]. All three features are regulatory genes with known functional roles in cancer metastasis, and their interactions suggest potential mechanisms of metastatic phenotype amplification (Table 5).

Discussion

In this study, we explored the extent to which transcriptomics and proteomics are informative of metastasis. Numerous studies have looked at the association between molecular profiles and cancer metastasis[14]. However, here, we use the combination of transcriptomics and proteomics, to build interpretable linear models predicting cancer metastasis in a continuous manner, and in a manner that generalizes to multiple tumor types.

First, we observed that linear machine learning models can predict metastasis as well as more complex non-linear models, including ANNs. Given their higher interpretability, it is beneficial to identify features predictive of metastasis with linear models for metastatic studies of similar sample sizes. However, this speaks in part to the large variance and moderate predictive power of these data in metastasis, yielding what appears to be an upper limit to the information standard machine learning approaches can extract. Additional omics layers, such as DNA-methylation[16], and additional samples with more variance, which may be limited in cancer cell lines, may help increase predictive performance. Thus, despite negligible discrepancies or poorer performance of non-linear models in the power analysis (Fig. 1b), it is possible that ANNs in particular would have better performance at much larger sample sizes than available here. Generally, more parameters require more samples in ANNs[29], while larger ANN models tend to perform better because of the double-descent behavior[30,31].

Next, we observed that transcriptomics predictions tend to outperform proteomic ones; this can in part be explained by the measurement of much fewer genes in proteomics relative to RNA-sequencing. Yet, when using the same starting set of genes, proteomics did not outperform transcriptomics, despite proteins typically being considered more proximal to mechanism or phenotype than mRNA. Notably, the proteomics pipeline typically selected substantially fewer features, indicating that when the relevant features are measured,

proteomics can get more information from those features as compared to transcriptomics. Additionally, each sequencing technology has distinct limitations; consequently, other contributing factors may include limits-of-detection preventing quantification of low-abundance proteins, which is not as much of an issue in RNA-seq[32] and higher noise in proteomics measurements[33].

Combining proteomics and transcriptomics did offer an increase in predictive performance as measured by Pearson correlation. The model fits combined features from both modalities in a non-redundant manner: there was significant depletion in the modality overlap of gene products in the top 500 genes and those that did overlap had relatively lower correlation. This indicates that a subset of proteomic features is more informative than their redundant transcriptomic counterpart (those that don't overlap) and another subset offers information complementary to that available in transcriptomics (those that do overlap). Furthermore, this indicates that across cell lines, the model fit is accounting for gene-specific variance in correlation[24] between mRNA and protein. The presence of non-linear relationships and network effects[34] between mRNA and protein can explain instances in which the two modalities are distinctly informative[20] of metastasis. We note that our observations regarding discrepancies in model performance between proteomic and transcriptomic modalities may not necessarily generalize to other datasets or phenotypes.

Finally, we observed that the most informative features in our linear joint omics model were associated with relevant processes such as cell migration and chemotaxis. This indicates that, despite moderate predictive performance, the model is learning biologically meaningful connections between metastasis and the molecular profile of the cell lines. We also found that, outside these top-ranked features, we could select for feature modules that participated in interactions that have substantial non-additive effects on metastatic potential, and that these features tended to have higher centrality in PPI networks.

Several genes with well-established roles in metastasis—SP1, APC, and AXIN1—recurrently exhibited strong effects in our analysis (Fig. 4b, Table 5). Notably, APC and AXIN1 both scaffold the β-catenin destruction complex, suggesting that convergent alterations of canonical Wnt signaling may have an outsized and generalizable impact on metastatic potential[28]. APC and AXIN1 are typically considered to act as tumor suppressors by downregulating canonical Wnt signaling, so their synergistic effect on metastatic potential may appear counterintuitive. However, over-expression of either component or alteration of their stoichiometric ratios modulate destruction complex stability and activity[35], and various mutations in either component, commonly found in DepMap cell lines, can instead stabilize β-catenin[36]. Both genes are also multi-functional, and their increased expression can contribute to metastasis in β-catenin independent manners; for example, APC plays a prominent role in reorganizing cytoskeletal components to enable cell migration [37] and over-expression of AXIN1 stimulates JNK signaling to trigger pro-metastatic programs[38–40]. PA2G4 showed antagonistic effects when paired with SP1, but synergistic effects with APC, potentially reflecting the distinct and sometimes opposing roles of its p42 (tumor-suppressive) and p48 (oncogenic) isoforms[41,42]. This highlights the importance of considering isoform-level information in the future. The recurring influence of

these features converges on major regulatory axes—including Wnt/β-catenin and MAPK/JNK—that control EMT, invasion, and immune evasion, highlighting key mechanistic entry points for metastatic progression.

These results are based on omics measurements of cancer cell lines, which might have substantial differences from *in vivo* tumors. *In vivo* metastasized tumors contain heterogeneous microenvironments influenced by mutational burden, interactions with tissue-specific cell types and the immune system, and nutrient availability and hypoxia[8,43]. Similarly, the metastatic potential was quantified in mice, and therefore might not directly translate to human metastatic phenotypes due to factors such as functional differences in orthologs[44], distinct global responses to perturbation[45], and pathogenesis through disease mechanisms that are subdominant in one species[46]. While molecular processes in these simpler models do not fully reflect *in vivo* responses, our pathway enrichment results indicate that the model still captures biologically relevant features. Nonetheless, tools[4547] that can contextualize and translate the omics profiles of cancer preclinical models to clinical responses can help improve predictive power, interpretability, and clinical relevance. Similarly, while our model is tissue-agnostic, explicitly accounting for the tissue-specific metastatic potential available in MetMap may improve specificity and predictive power. Altogether, these findings highlight the value of multi-omics data integration in interpretable models, not only to enhance predictive accuracy but also to reveal how features from different modalities jointly associate with a phenotype of interest.

Methods

Transcriptomic and Proteomic Data Processing

Metastatic potential for 488 samples reported by MetMap 500[19] was downloaded from https://depmap.org/metmap/data/index.html.

Batch-corrected RNA-sequencing data of protein-coding genes for the DepMap[17] cell lines was downloaded in log-normalized transcripts per million (https://depmap.org/portal/data_page/?tab=allData&releasename=DepMap%20Public%2024Q4&filename=OmicsExpressionProteinCodingGenesTPMLogp1BatchCorrected.csv). 481 samples overlapped out of those reported in MetMap 500 and the 1673 reported in the RNA-sequencing dataset.

Normalized protein expression data reported in Table S2 of Nusinow et al.[18] were used. Proteins with missing values in more than 80% of samples (1,786 of 12,755 features) were removed. The remaining missing values were imputed in Perseus (version 2.1.3.0) as follows: the protein distribution was transformed by 1) subtracting 1.8 standard deviations from all values and then 2) multiplying all values by 0.3 to shrink the standard deviation; next, imputed values were drawn by random sampling from this transformed distribution. Three cell lines had two samples originating from different ten-plexes, which were dropped. Finally, 248 samples overlapped out of those reported in MetMap 500 and the 372 reported in the proteomics dataset.

247 samples were present in both the transcriptomics and proteomics data.

Model Fitting and Hyperparameter Tuning

The following models were each fit on the data, with omics as the input and metastatic potential as the output: partial least squared (PLS) regression, Ridge, Lasso, ElasticNet, Support Vector Machine (SVM) with a linear kernel. Similarly, the following nonlinear regression models were fit on the data: SVM with polynomial and RBF kernels, k-Nearest Neighbors (KNN), and Random Forest.

Models were either fit on transcriptomics separately, proteomics separately, or the two datasets jointly. The following preprocessing steps were always conducted on either modality separately: the top features were selected according to those that had the highest coefficient of variation (feature selection is only conducted on training data), and proceeded by mean-centering. The number of features used by the model was one of the hyperparameters explored (see Table 3 for a comprehensive accounting of hyperparameters explored).

Each model type was fit on the same 10 train-test splits using 10-fold CV. Within each fold, an inner 5-fold CV on the training data was used to identify the best set of hyperparameters (Table 3, sheet 1). Hyperparameter tuning was conducted using Optuna[48] with 100 trials to minimize the average mean squared error. The “SuccessiveHalvingPruner” pruner was used alongside a custom sampler. The custom sampler used the CmaEsSampler on hyperparameters when possible, and the TPESampler otherwise. To prevent convergence to a local minimum, every 20 trials, hyperparameters were also randomly sampled using the TPESampler.

Upon identification of the best set of hyperparameters, model predictions were assessed on the outer fold test data using Pearson correlation and mean squared error (Fig. S1). This resulted in a unique set of hyperparameters for each fold for each model type (Table 3, sheets 2-4). A single consensus model for each of the top three performing model types (ranked by mean performance across the 10 folds) for both performance metrics was generated as follows: 1) for discrete and integer hyperparameters, the mode across 10-folds was used, and 2) for continuous hyperparameters, the mean value across the 10-folds was used. The exception here was the number of features selected for proteins in the joint omics pipeline, in which the median rather than the mode was used because there were more uniform distributions across folds. Once a consensus model was identified, it was assessed using the same performance metrics on a different set of folds using 10-fold CV (Fig. 1A, Fig. S4A, and Fig. S5). 10-fold CV was also used to assess single and ensemble artificial feed-forward neural networks (ANNs) (see Supplementary Methods for details).

Power Analysis of Transcriptomic Model

To assess how model performance scales with training sample size, we conducted a power analysis using the top-performing consensus models fit on transcriptomics. We trained models on 10–100% of the training data in 10% increments using 10-fold CV, repeating each subsample 100 times to generate stable estimates. Performance was evaluated using mean squared error

(MSE) and Pearson correlation, the latter Fisher z-transformed prior to modeling. We compared linear, power, and exponential fits to describe the relationship between sample size and model performance, selecting the best function based on adjusted R^2, AIC, and BIC. Power functions best explained the performance trends across models and metrics. We then fit a power regression model with model type as a covariate to quantify performance differences across sample sizes. For additional details, see the Supplementary Methods.

Feature Interactions

To screen for non-additive effects we began with the consensus joint-omics linear SVR and introduced an interaction term (defined as the sample-wise product between the two features' expression values) evaluated every candidate gene-pair in two complementary ways.

- **Approach A:** We regressed SVR residuals on the interaction term and assigned significance from the OLS t-statistic.

- **Approach B:** We refit the SVR with the added interaction term, compared the resulting coefficient to a 1,000-fold permutation null, and derived a two-sided empirical p-value.

Because an exhaustive search over the selected 19,638 features (>1.9e8 pairs) is intractable, we deployed a genetic algorithm (GA) that evolves *sets* of 25 features. Focusing on feature sets allows us to capture recurring interaction modules, thereby identifying features that participate in multiple interactions and may be central to synergistic or antagonistic modules. The GA maximised a weighted sum of effect size ($|\beta|$) and signal-to-noise ($|\beta|/SE$) from Approach A, was initialised with 720 random individuals, and ran ≤ 300 generations with crossover, mutation (10%), periodic re-initialisation, and early-stopping safeguards. We executed the GA on (i) all features, (ii) The Cancer Gene Census[49] genes, and (iii) Cancer Cell Map[50] genes, yielding three candidate sets (900 total interactions).

Interactions passing BH-FDR thresholds (≤0.20 for the search on all genes, ≤0.15 for the search on database genes) were re-tested with Approach B. Those passing an FDR ≤ 0.20 were deemed high-confidence interactions.

For each high-confidence pair, we conducted *in-silico* perturbations. While holding all other genes at an empirically derived control vector, we uniformly sampled values for the two partners across their observed expression ranges and compared predictions from the SVR with and without the interaction term. The resulting difference surface classified synergistic ($\Delta > 0$) versus antagonistic ($\Delta < 0$) regions while remaining interpolative, ensuring reliable estimates of the interactions' impact on metastatic potential. For additional details on feature interaction analysis, see Supplementary Methods.

References

1. Guan, X. Cancer metastases: challenges and opportunities. *Acta Pharm Sin B* **5**, 402–418 (2015).
2. Chaffer, C. L. & Weinberg, R. A. A perspective on cancer cell metastasis. *Science* **331**, 1559–1564 (2011).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
4. D'Orazio, M. *et al.* Deciphering cancer cell behavior from motility and shape features: Peer prediction and dynamic selection to support cancer diagnosis and therapy. *Front. Oncol.* **10**, 580698 (2020).
5. Stuelten, C. H., Parent, C. A. & Montell, D. J. Cell motility in cancer invasion and metastasis: insights from simple model organisms. *Nat. Rev. Cancer* **18**, 296–312 (2018).
6. Shi, X. *et al.* Mechanism insights and therapeutic intervention of tumor metastasis: latest developments and perspectives. *Signal Transduct. Target. Ther.* **9**, 192 (2024).
7. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
8. Baghdassarian, H. M. & Lewis, N. E. Resource allocation in mammalian systems. *Biotechnol Adv* **71**, 108305 (2024).
9. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* **10**, 291–305 (2012).
10. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annu Rev Cell Dev Biol* **26**, 721–744 (2010).
11. Subramanian, A. *et al.* A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
12. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv* **49**, 107739 (2021).
13. Zhang, Z. *et al.* Morphology-based prediction of cancer cell migration using an artificial neural network and a random decision forest. *Integr. Biol.* **10**, 758–767 (2018).
14. Albaradei, S. *et al.* Machine learning and deep learning methods that use omics data for metastasis prediction. *Comput. Struct. Biotechnol. J.* **19**, 5008–5018 (2021).
15. Lee, S. C., Quinn, A., Nguyen, T., Venkatesh, S. & Quinn, T. P. A cross-cancer metastasis signature in the microRNA-mRNA axis of paired tissue samples. *Mol. Biol. Rep.* **46**, 5919–5930 (2019).
16. Albaradei, S. *et al.* MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput. Struct. Biotechnol. J.* **19**, 4404–4411 (2021).
17. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
18. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387–402.e16 (2020).
19. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
20. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**, 630–644 (2020).
21. Yamaguchi, H., Wyckoff, J. & Condeelis, J. Cell migration in tumors. *Curr. Opin. Cell Biol.* **17**, 559–564 (2005).
22. Hood, J. D. & Cheresh, D. A. Role of integrins in cell invasion and migration. *Nat. Rev. Cancer* **2**, 91–100 (2002).
23. Tochhawng, L., Deng, S., Pervaiz, S. & Yap, C. T. Redox regulation of cancer cell migration and invasion. *Mitochondrion* **13**, 246–253 (2013).

24. Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12**, 883 (2016).
25. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
26. Yang, Y., Zhang, S. & Guo, L. Characterization of Cell Cycle-Related Competing Endogenous RNAs Using Robust Rank Aggregation as Prognostic Biomarker in Lung Adenocarcinoma. *Front Oncol* **12**, 807367 (2022).
27. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol* **17**, e9923 (2021).
28. Liu, J. *et al.* Wnt/β-catenin signalling: function, biological mechanisms, and therapeutic opportunities. *Signal Transduct Target Ther* **7**, 3 (2022).
29. Alwosheel, A., van Cranenburgh, S. & Chorus, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J. Choice Model.* **28**, 167–182 (2018).
30. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15849–15854 (2019).
31. M., C. & Hugh. *Deep Learning: Foundations and Concepts*. (Springer International Publishing, Cham, Switzerland, 2023).
32. Wang, X., Liu, Q. & Zhang, B. Leveraging the complementary nature of RNA-Seq and shotgun proteomics data. *Proteomics* **14**, 2676–2687 (2014).
33. De Los Santos, H., Bennett, K. P. & Hurley, J. M. MOSAIC: a joint modeling methodology for combined circadian and non-circadian analysis of multi-omics data. *Bioinformatics* **37**, 767–774 (2021).
34. Sweatt, A. J. *et al.* Proteome-wide copy-number estimation from transcriptomics. *Molecular Systems Biology* **20**, 1230–1256 (2024).
35. Schaefer, K. N. *et al.* Supramolecular assembly of the beta-catenin destruction complex and the effect of Wnt signaling on its localization, molecular size, and activity in vivo. *PLoS Genet* **14**, e1007339 (2018).
36. Stamos, J. L. & Weis, W. I. The β-catenin destruction complex. *Cold Spring Harb Perspect Biol* **5**, a007898 (2013).
37. Fang, X. & Svitkina, T. M. Adenomatous Polyposis Coli (APC) in cell migration. *Eur J Cell Biol* **101**, 151228 (2022).
38. Zhang, Y., Neo, S. Y., Wang, X., Han, J. & Lin, S. C. Axin forms a complex with MEKK1 and activates c-Jun NH(2)-terminal kinase/stress-activated protein kinase through domains distinct from Wnt signaling. *J Biol Chem* **274**, 35247–35254 (1999).
39. Qiu, L. *et al.* The scaffold protein AXIN1: gene ontology, signal network, and physiological function. *Cell Commun Signal* **22**, 77 (2024).
40. Ebelt, N. D., Cantrell, M. A. & Van Den Berg, C. L. c-Jun N-Terminal Kinases Mediate a Wide Range of Targets in the Metastatic Cascade. *Genes Cancer* **4**, 378–387 (2013).
41. Jia, W. *et al.* PA2G4 in health and disease: An underestimated multifunctional regulator. *J Adv Res* (2025) doi:10.1016/j.jare.2025.02.002.
42. Stevenson, B. W. *et al.* A structural view of PA2G4 isoforms with opposing functions in cancer. *J Biol Chem* **295**, 16100–16112 (2020).
43. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med* **24**, 541–550 (2018).
44. Liao, B.-Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**, 6987–6992 (2008).
45. Meimetis, N. *et al.* AutoTransOP: translating omics signatures without orthologue requirements using deep learning. *NPJ Syst. Biol. Appl.* **10**, 13 (2024).
46. Pullen, K. M. *et al.* Cross species modeling reveals a role for the unfolded protein response

in shaping the transcriptional reaction to Mycobacterium tuberculosis infection. *bioRxiv* 2024.04.19.590290 (2024) doi:10.1101/2024.04.19.590290.
47. Dimitrieva, S. *et al.* Biologically relevant integration of transcriptomics profiles from cancer cell lines, patient-derived xenografts, and clinical tumors using deep learning. *Sci Adv* **11**, eadn5596 (2025).
48. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. (2019).
49. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696–705 (2018).
50. Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A. & Ideker, T. The cancer cell map initiative: defining the hallmark networks of cancer. *Mol Cell* **58**, 690–698 (2015).
51. O'Connor, L., Gilmour, J. & Bonifer, C. The Role of the Ubiquitously Expressed Transcription Factor Sp1 in Tissue-specific Transcriptional Regulation and in Disease. *Yale J Biol Med* **89**, 513–525 (2016).
52. Yoo, J. Y. *et al.* Interaction of the PA2G4 (EBP1) protein with ErbB-3 and regulation of this binding by heregulin. *Br J Cancer* **82**, 683–690 (2000).
53. Kim, J. Y. *et al.* β-arrestin 2 negatively regulates lung cancer progression by inhibiting the TRAF6 signaling axis for NF-κB activation and autophagy induced by TLR3 and TLR4. *Cell Death Dis* **14**, 422 (2023).
54. Li, L. *et al.* ARRB2 promotes cervical cancer progression via stabilizing CDC25A mRNA through m6A-IGF2BP1-dependent manner. *NPJ Precis Oncol* **9**, 80 (2025).
55. Mathews, P. *et al.* β-Arrestin 2 as a Prognostic Indicator and Immunomodulatory Factor in Multiple Myeloma. *Cells* **14**, (2025).
56. Zhou, B. *et al.* The Comprehensive Analysis of Hub Gene ARRB2 in Prostate Cancer. *Dis Markers* **2022**, 8518378 (2022).
57. Fodde, R. The APC gene in colorectal cancer. *Eur J Cancer* **38**, 867–871 (2002).
58. Zheng, Y. *et al.* The RNA-binding protein PCBP1 represses lung adenocarcinoma progression by stabilizing DKK1 mRNA and subsequently downregulating β-catenin. *J Transl Med* **20**, 343 (2022).
59. Huang, S. *et al.* PCBP1 regulates the transcription and alternative splicing of metastasis-related genes and pathways in hepatocellular carcinoma. *Sci Rep* **11**, 23356 (2021).
60. Lee, J., You, J. H. & Roh, J.-L. Poly(rC)-binding protein 1 represses ferritinophagy-mediated ferroptosis in head and neck cancer. *Redox Biol* **51**, 102276 (2022).
61. Feng, P. *et al.* NONO and tumorigenesis: More than splicing. *J Cell Mol Med* **24**, 4368–4376 (2020).
62. Kais, Z. & Parvin, J. D. Regulation of centrosomes by the BRCA1-dependent ubiquitin ligase. *Cancer Biol Ther* **7**, 1540–1543 (2008).
63. Li, H., Sekine, M., Tung, N. & Avraham, H. K. Wild-type BRCA1, but not mutated BRCA1, regulates the expression of the nuclear form of beta-catenin. *Mol Cancer Res* **8**, 407–420 (2010).
64. Huen, M. S. Y., Sy, S. M. H. & Chen, J. BRCA1 and its toolbox for the maintenance of genome integrity. *Nat Rev Mol Cell Biol* **11**, 138–148 (2010).
65. Sakamoto, A. *et al.* Cross-talk between the transcription factor Sp1 and C/EBPβ modulates TGFβ1 production to negatively regulate the expression of chemokine RANTES. *Heliyon* **4**, e00679 (2018).
66. Hwang, I., Ko, H. R. & Ahn, J.-Y. The roles of multifunctional protein ErbB3 binding protein 1 (EBP1) isoforms from development to disease. *Exp Mol Med* **52**, 1039–1047 (2020).
67. Biddlestone, J., Batie, M., Bandarra, D., Munoz, I. & Rocha, S. SINHCAF/FAM60A and SIN3A specifically repress HIF-2α expression. *Biochem J* **475**, 2073–2090 (2018).
68. Alroy, I., Soussan, L., Seger, R. & Yarden, Y. Neu differentiation factor stimulates

phosphorylation and activation of the Sp1 transcription factor. *Mol Cell Biol* **19**, 1961–1972 (1999).
69. Xiong, S. *et al.* Up-regulation of vascular endothelial growth factor in breast cancer cells by the heregulin-beta1-activated p38 signaling pathway enhances endothelial cell migration. *Cancer Res* **61**, 1727–1732 (2001).
70. Müller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res* **51**, 10934–10949 (2023).
71. Kallifatidis, G., Mamouni, K. & Lokeshwar, B. L. The Role of β-Arrestins in Regulating Stem Cell Phenotypes in Normal and Tumorigenic Cells. *Int J Mol Sci* **21**, (2020).
72. Tohgo, A., Pierce, K. L., Choy, E. W., Lefkowitz, R. J. & Luttrell, L. M. beta-Arrestin scaffolding of the ERK cascade enhances cytosolic ERK activity but inhibits ERK-mediated transcription following angiotensin AT1a receptor stimulation. *J Biol Chem* **277**, 9429–9436 (2002).
73. Chu, S. & Ferro, T. J. Sp1: regulation of gene expression by phosphorylation. *Gene* **348**, 1–11 (2005).
74. Beaulieu, J.-M. *et al.* An Akt/beta-arrestin 2/PP2A signaling complex mediates dopaminergic neurotransmission and behavior. *Cell* **122**, 261–273 (2005).
75. Horstkorte, R. *et al.* Selective inhibition of polysialyltransferase ST8SiaII by unnatural sialic acids. *Exp Cell Res* **298**, 268–274 (2004).
76. Peng, K., Chen, X., Lin, A., Tong, Z. & Lin, W. PolyC-RNA-binding protein 1 (PCBP1) enhances tropomyosin 3 (TPM3) mRNA stability to promote the progression of esophageal squamous cell carcinoma. *Bioengineered* **13**, 8581–8592 (2022).
77. Fife, C. M., McCarroll, J. A. & Kavallaris, M. Movers and shakers: cell cytoskeleton in cancer metastasis. *Br J Pharmacol* **171**, 5507–5523 (2014).
78. Catalano, T. *et al.* Oxidative Distress Induces Wnt/β-Catenin Pathway Modulation in Colorectal Cancer Cells: Perspectives on APC Retained Functions. *Cancers (Basel)* **13**, (2021).
79. Lone, B. A. *et al.* Non-POU Domain-Containing Octomer-Binding (NONO) protein expression and stability promotes the tumorigenicity and activation of Akt/MAPK/β-catenin pathways in human breast cancer cells. *Cell Commun Signal* **21**, 157 (2023).
80. Peng, Y. *et al.* A novel protein AXIN1-295aa encoded by circAXIN1 activates the Wnt/β-catenin signaling pathway to promote gastric cancer progression. *Mol Cancer* **20**, 158 (2021).
81. Sun, S. *et al.* PA2G4 promotes the metastasis of hepatocellular carcinoma by stabilizing FYN mRNA in a YTHDF2-dependent manner. *Cell Biosci* **12**, 55 (2022).
82. Li, K. *et al.* Tyrosine kinase Fyn promotes osteoarthritis by activating the β-catenin pathway. *Ann Rheum Dis* **77**, 935–943 (2018).
83. Bao, Y. *et al.* Ebp1 p48 promotes oncogenic properties in hepatocellular carcinoma through p38 MAPK/HIF1α activation and p53 downregulation. *Mol Carcinog* **60**, 252–264 (2021).
84. Powell, E., Piwnica-Worms, D. & Piwnica-Worms, H. Contribution of p53 to metastasis. *Cancer Discov* **4**, 405–414 (2014).
85. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. in *Advances in Neural Information Processing Systems* 6402–6413.

Acknowledgements

We would like to thank Luka Karginov and Brian Joughin for helpful discussions regarding correlation networks and feature selection, respectively. We would like to thank Douglas

Lauffenburger and Nathan Lewis for helpful discussions regarding the relationships between transcriptomics and proteomics. H.M.B is funded as a MIT-Novo Nordisk Artificial Intelligence Postdoctoral Fellow. N.M. is a 2024-2025 Takeda Fellow.

Ethics Declaration

Declaration of Interests

The authors declare no competing interests.

Data Availability

All data is publicly available through DepMap (https://depmap.org/portal/) and CCLE (https://sites.broadinstitute.org/ccle/).

Code Availability

Code for all analyses in this manuscript are available on Github https://github.com/hmbaghdassarian/metastatic_potential.